A general technique to train language models on language models

Nederhof, MJ

Published in:
Computational Linguistics

DOI:
10.1162/0891201054223986

Publication date:
2005

Book Reviews

A Computational Model of Natural Language Communication

Roland Hausser
(Friedrich-Alexander-Universität Erlangen-Nürnberg)

Springer, 2006, xii+365 pp; hardbound, ISBN 3-540-35476-X/978-3-540-35476-5, €69.50

Reviewed by
Markus Egg
University of Groningen

The work presented in this book is motivated by the goal of applying linguistic theory-building to the concrete needs of potential linguistic applications such as question answering, dialogue systems, and machine translation. To pursue this goal, a translation of linguistic theory into a framework of "practical linguistics" is suggested. *Database Semantics* (DBS) is presented as a first step towards such a framework. It models the communication between cognitive agents, which can be used, for example, to implement the communicative abilities of a cognitive robot.

DBS serves as a single underlying format for modeling communication in that it lends itself to an account of both language processing and language production (thinking is added as a separate component, which refers to inferencing on stored information, and activating content to be verbalized). As such an underlying format, it can be used to describe linguistic as well as extralinguistic content (to represent utterances and the context, respectively). Being explicitly designed for practical applications, DBS deliberately ignores linguistic phenomena considered irrelevant for these (e.g., quantifier scope).

The structure of the book is as follows. It has three main parts, which introduce DBS, outline the range of constructions covered by DBS so far, and specify fragments that can be processed or produced in the framework of DBS. There is also an appendix with two sections on the treatment of word-order variation in DBS and on the global architecture of DBS systems, and a glossary.

The first part of the book starts with general principles of linguistic analysis that apply to DBS. These principles include *incrementality* (input is to be processed successively as it comes in, which yields an analysis for incomplete as well as complete chunks of input; the syntactic basis for this strategy is Left-Associative Grammar [Hausser 1992]), *surface orientation* (no empty categories), and a focus on *communication* (description formalisms must be able to handle turn-taking, i.e., language processing and production).

After a sketch of the general theory of communication of which DBS is a part, DBS is presented in detail. It is implemented as a non-recursive data structure, that is, a list of feature structures called **proplets** (usually, one per word[1]) that are linked by coindexing the values of specific features.[2] For example, subcategorizing elements ("functors") have features whose values indicate their arguments and the other way around.

In spite of its name, DBS does not offer a purely semantic representation of linguistic expressions. Although it does abstract away from purely syntactic phenomena such

1 Function words such as determiners, auxiliaries, and conjunctions have no proplets of their own but contribute to other proplets.
2 This technique makes it resemble minimal recursion semantics (Copestake et al. 2005).

as word order and diatheses, it still preserves much syntactic structure, for example, in its representation of modification and of elliptical expressions. Semantics proper is encoded within proplets (except those for deictic expressions and proper names) by defining a concept as the value of their "core attribute."

DBS also serves for the representation of the extralinguistic context. The context is described in terms of proplet sets that are linked by feature value coindexation; the only difference to proplet sets for the modeling of linguistic content is that proplet sets for context do not comprise explicit pointers to specific words.

The similarity between the representations of utterances and of context makes the move between them straightforward, which is crucial for the proposed analysis of language processing and production: Language processing consists of deriving lists of proplets (including the coindexations between proplet values) from utterances and storing them in the context representation, which is modeled as a database. Language production consists of the activation of such lists of proplets from this database and their translation into utterances.

The second part of the book is devoted to three classes of linguistic phenomena and their description in DBS. The first class is called "functor-argument structure" and covers the relations between subcategorizing elements and their arguments and modification. This includes sentential arguments, subordinate clauses, and relative clauses. The second class consists of coordination phenomena, ranging from simple coordination on the word or phrase level to gapping and right-node raising. The last class is cases of coreference. A wide range of these cases is represented in DBS, including even Bach–Peters sentences (where there are two NPs that constitute anaphors whose antecedent is the respective other NP). The DBS framework is used to formulate a version of the Langacker–Ross condition dating back to Langacker (1969) and Ross (1969): Pronouns can precede a coreferential NP only if they are part of a clause that is embedded within the clause of the NP.

In the third part, three fragments are presented in detail, the first two from the processing and production perspective, the last one only from the processing perspective. The first fragment prepares the ground by illustrating how the approach handles extremely simple texts consisting of intransitive present-tense sentences whose NP is a proper name. The second fragment extends the coverage to pronouns, complex NPs (Det-Adj*-N), and transitive and ditransitive verbs in simple and complex tenses. Finally, the third fragment offers a treatment of intensifiers (*very, rather*) and adverbials, and an outlook on a syntactically underspecified approach to modifier attachment ambiguities. The fragments are described in terms of "grammars," which specify start and end states (in terms of the first and the last proplet of a list to be processed or verbalized) and a set of rules. The rules are ordered in that every rule is accompanied by a set of potential successors, and in that rules to start and to end a derivation with are specified.

The book is written in a highly accessible way. The formalism itself as well as its application to the fragments is described thoroughly, which makes it easy to understand and evaluate DBS. The underlying perspective on linguistic theory-building and the theory of communication of which DBS is a part are also explicated clearly. The formal details of the analysis are presented carefully. A remaining point of dispute is in my view the set of readings of sentences where several PPs have more than one attachment possibility (Chapter 15.1).

However, the book does not offer much discussion of the relation between the proposed analysis and competing approaches. This shows up in specific parts of the analysis—for example, in the discussion of coreference in Chapter 10, which does not

integrate previous work that formulates constraints on potential coreferences in terms of syntactic constellations such as *c*- or *o*-command (e.g., Pollard and Sag 1994; Reuland 2006), and in the treatment of quantifier scope and scope ambiguity in Chapter 6 (as opposed to, e.g., the papers in van Deemter and Peters [1996]). But even more important, it would have been interesting to hear more about the way in which DBS compares to other approaches whose goal is the application of linguistic theory-building to concrete needs of potential linguistic applications. Although the completion of the manuscript admittedly antedates much of the ongoing work in the field (e.g., the application of deep linguistic processing in the analysis of biomedical and other scientific texts), a comparison of DBS to wide-coverage systems such as the LinGO English Resource Grammar (Copestake and Flickinger 2000) (including also related activities such as the development of Robust Minimal Recursion Semantics [Copestake 2007]) or Alpino (analysis of unrestricted Dutch texts [Bouma, van Noord, and Malouf 2001]) would have been a welcome complementation to the presentation of DBS in the book.

References

Bouma, Gosse, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide coverage computational analysis of Dutch. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands (CLIN) 2000*, Rodopi, Amsterdam, pages 45–59.

Copestake, Ann. 2007. Semantic composition with (robust) minimal recursion semantics. In *Proceedings of the Workshop on Deep Processing*, pages 73–80, Prague.

Copestake, Ann and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens.

Copestake, Ann, Daniel Flickinger, Carl Pollard, and Ivan Sag. 2005. Minimal recursion semantics. An introduction. *Research on Language and Computation*, 3:281–332.

van Deemter, Kees and Stanley Peters, editors. 1996. *Semantic Ambiguity and Underspecification*. CSLI, Stanford, CA.

Hausser, Roland. 1992. Complexity in left-associative grammar. *Theoretical Computer Science*, 106:283–308.

Langacker, R. 1969. On pronominalization and the chain of command. In D. Reibel and S. Schane, editors, *Modern Studies in English*. Prentice Hall, Englewood Cliffs, NJ, pages 160–186.

Pollard, Carl and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. CSLI and University of Chicago Press, Chicago, IL.

Reuland, Eric. 2006. Binding theory: Terms and concepts. In Martin Everaert and Henk van Riemsdijk, editors, *The Blackwell Companion to Syntax*, volume 1. Blackwell, Malden, UK, chapter 9, pages 260–283.

Ross, J. 1969. On the cyclic nature of English pronominalization. In D. Reibel and S. Schane, editors, *Modern Studies in English*. Prentice Hall, Englewood Cliffs, NJ, pages 187–200.

Markus Egg is an associate professor in Discourse Studies at the University of Groningen. His main areas of interest are semantics and discourse and their interfaces with syntax. His address is Centre for Language and Cognition Groningen, Rijksuniversiteit Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands; e-mail: k.m.m.egg@rug.nl.